**Multiple Regression Analysis**

A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working (Salary), their age (Age), how many years they had worked as a model (Years), and then got a panel of experts from modeling agencies to rate the attractiveness of each model as a percentage, with 100% being perfectly attractive (Beauty). I ran a multiple regression model to predict the salary of models using age, years and beauty.

**Were these met? Why or why not?**

The assumptions are as follows:

**Linearity -** the relationships between the predictors and the outcome variable should be linear

**Normality** - the residuals should be normally distributed

**Homogeneity of variance (homoscedasticity)** - the residual variance should be constant

**Independence -** the residuals associated with one observation are not correlated with the errors of any other observation

**Citations?**

| **Correlations** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | Salary per Day (£) | Age (Years) | Number of Years as a Model | Attractiveness (%) |
| Salary per Day (£) | Pearson Correlation | 1 | .397\*\* | .337\*\* | .068 |
| Sig. (2-tailed) |  | .000 | .000 | .304 |
| N | 231 | 231 | 231 | 231 |
| Age (Years) | Pearson Correlation | .397\*\* | 1 | .955\*\* | .261\*\* |
| Sig. (2-tailed) | .000 |  | .000 | .000 |
| N | 231 | 231 | 231 | 231 |
| Number of Years as a Model | Pearson Correlation | .337\*\* | .955\*\* | 1 | .173\*\* |
| Sig. (2-tailed) | .000 | .000 |  | .008 |
| N | 231 | 231 | 231 | 231 |
| Attractiveness (%) | Pearson Correlation | .068 | .261\*\* | .173\*\* | 1 |
| Sig. (2-tailed) | .304 | .000 | .008 |  |
| N | 231 | 231 | 231 | 231 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | |

From the correlation matrix it is clear that there is very strong correlation exists between Age and Number of years. Hence there might be a chance of multicollinearity. Possible alternatives are:

If appropriate combine these two variables or

Remove one of the variables from the model



From the histogram it is clear that the distribution of residuals is positively skewed. Hence the normality of residuals is not satisfied.



Since there is some pattern for the points, the points on the plot of residuals against the fitted value are not at random. Hence we can conclude that the errors are dependent and the residual variances are not constant. Possible alternatives are:

Transform the data so regression is appropriate

**Write out if the assumptions are met in a paragraph (see stats assignment template) so I don’t have to search for it among data output**

Weighted least squares regression

**Syntax:**

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT salary

/METHOD=ENTER age years beauty

/SCATTERPLOT=(\*ZRESID ,\*ZPRED)

/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).

CORRELATIONS

/VARIABLES=salary age years beauty

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE.

The estimated regression equation is given by,

Salary = -60.89 + Age \* 6.234 – Years \* 5.561 – Beauty \* 0.196

| **Coefficientsa** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -60.890 | 16.497 |  | -3.691 | .000 |
| Age (Years) | 6.234 | 1.411 | .942 | 4.418 | .000 |
| Number of Years as a Model | -5.561 | 2.122 | -.548 | -2.621 | .009 |
| Attractiveness (%) | -.196 | .152 | -.083 | -1.289 | .199 |
| a. Dependent Variable: Salary per Day (£) | | | | | | |

The regression coefficients can be interpreted as follows:

For a unit increase in Age, the salary increases by 6.234 units.

For a unit increase in years, the salary decreases by 5.561 units.

For a unit increase in beauty rating (%), the salary decreases by 0.196 units.

The significance of the regression coefficients are tested using Student’s t test.

**Age**

The **null hypothesis** tested is

H0: Age is not significant in predicting the salary of models (β1 = 0) **What about the other variables?**

The **alternative hypothesis** is

H1: Age is significant in predicting the salary of models (β1 ≠ 0)

Test statistic:  = 4.418

**Decision rule:** Reject the null hypothesis if the observed significance (P-value) is less than the significance level 0.05.

P value = 0.0000

**Conclusion:** Reject the null hypothesis, since the observed significance (P value) is less than the significance level 0.05. The data provides enough evidence to conclude that age is significant in predicting the salary of models.

**Years**

The **null hypothesis** tested is

H0: Years are not significant in predicting the salary of models (β2 = 0)

The **alternative hypothesis** is

H1: Years are significant in predicting the salary of models (β2 ≠ 0)

Test statistic: = -2.621 **Can put all variables in one null hypothesis statement**

**Decision rule:** Reject the null hypothesis if the observed significance (P-value) is less than the significance level 0.05.

P value = 0.009

**Conclusion:** Reject the null hypothesis, since the observed significance (P value) is less than the significance level 0.05. The data provides enough evidence to conclude that years are significant in predicting the salary of models.

**Beauty**

The **null hypothesis** tested is

H0: Beauty is not significant in predicting the salary of models (β3 = 0)

The **alternative hypothesis** is

H1: Beauty is significant in predicting the salary of models (β3 ≠ 0)

Test statistic:  = -1.289 **Put all together**

**Decision rule:** Reject the null hypothesis if the observed significance (P-value) is less than the significance level 0.05.

P value = 0.199

**Conclusion:** Fails to reject the null hypothesis, since the observed significance (P value) is greater than the significance level 0.05. The data does not provide enough evidence to conclude that beauty is significant in predicting the salary of models.

The significance of the regression model is tested using F-test.

Here F-statistic = 17.066

P-value = 0.000

Since the F-statistic is significant with p-value less than 0.05, we can conclude that the estimated regression model is significant in predicting the salary of the models.

**Details**

| **ANOVAb** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 10871.964 | 3 | 3623.988 | 17.066 | .000a |
| Residual | 48202.790 | 227 | 212.347 |  |  |
| Total | 59074.754 | 230 |  |  |  |
| a. Predictors: (Constant), Attractiveness (%), Number of Years as a Model, Age (Years)  b. Dependent Variable: Salary per Day (£) | | | | | | |

The model adequacy is measured using the R2 value. Here R2 = 0.184. Thus 18.4% variability in the salary can be explained by the regression model. Thus the suggested regression model is not able to explain a fair portion of the variability in the dependent variable.

**Details**

| **Model Summaryb** | | | | |
| --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .429a | .184 | .173 | 14.57213 |
| a. Predictors: (Constant), Attractiveness (%), Number of Years as a Model, Age (Years)  b. Dependent Variable: Salary per Day (£) | | | | |

Reference: **Where were these used? I didn’t see citations. Not APA format**

Lewis-Beck MS, (1993). Regression Analysis, Beverley Hills, CA: Sage.

Wayne DW, (1995). Biostatistics, 6th ed, New York: John Wiley & Sons.

Draper NR, Smith H, (1981). Applied Regression Analysis. 2nd ed. New York: John Wiley & Sons.

Grading rubric:

Assumptions (5): 5

Assumptions met? (5): 5

Null and Alt Hyp (5): 4—put all variables in one statement

Syntax (5): 5

Output (5): 5

Results (10): 10

APA (10): 5—no citations, references not in APA format

Power and effect size (5): 0—not mentioned

Total (50): 39